(19) Europäisches Patentamt
European Patent Office
Office européen des brevets

(11) Publication number: **0 196 237 B1**

(12) # EUROPEAN PATENT SPECIFICATION

(45) Date of publication of patent specification: **17.06.92** (51) Int. Cl.[5]: **G06F 15/40**

(21) Application number: **86302323.0**

(22) Date of filing: **27.03.86**

The file contains technical information submitted after the application was filed and not included in this specification

(54) **Method of storing and searching chemical structure data.**

(30) Priority: **29.03.85 JP 63283/85**

(43) Date of publication of application:
01.10.86 Bulletin 86/40

(45) Publication of the grant of the patent:
17.06.92 Bulletin 92/25

(84) Designated Contracting States:
AT CH DE FR GB IT LI NL

(56) References cited:
EP-A- 0 090 895

JOURNAL OF CHEMICAL DOCUMENTATION, vol. 8, no. 2, May 1968, pages 113-122, American Chemical Society, Washington, D.C., US; N. JOCHELSON et al.: "The automation of structural group contribution methods in the estimation of physical properties"

IBM JOURNAL OF RESEARCH AND DEVELOPMENT, vol. 1, January 1964, pages 22-32, New York, US; R.E. BONNER: "On some clustering techniques"

(73) Proprietor: **Japan Association for International Chemical Information**
4-16, Yayoi 2-chome Bunkyo-ku
Tokyo 113(JP)

(72) Inventor: **Tokizane, Soichi**
c/o Japan Ass. for Int. Chemical Information
4-16, Yayoi 2-chome Bunkyo-ku Tokyo 113(JP)
Inventor: **Chihara, Hideaki**
c/o Japan Ass. for Int. Chemical Information
4-16, Yayoi 2-chome Bunkyo-ku Tokyo 113(JP)

(74) Representative: **Charlton, Peter John et al**
Elkington and Fife Prospect House 8 Pembroke Road
Sevenoaks, Kent TN13 1XR(GB)

Note: Within nine months from the publication of the mention of the grant of the European patent, any person may give notice to the European Patent Office of opposition to the European patent granted. Notice of opposition shall be filed in a written reasoned statement. It shall not be deemed to have been filed until the opposition fee has been paid (Art. 99(1) European patent convention).

EP 0 196 237 B1

JOURNAL OF CHEMICAL INFORMATION AND COMPUTER SCIENCES, vol. 24, no. 4, November 1984, pages 220-229, American Chemical Society, Washington, D.C., US; H. ABE et al.: "A computer program for generation of constitutionally isomeric structural formulas"

JOURNAL OF THE AMERICAN CHEMICAL SOCIETY, vol. 96, no. 15, 24th July 1974, pages 4825-4834, Gaston, US; W. WIPKE et al.: "Simulation and evaluation of chemical synthesis. Computer representation and manipulation of stereochemistry"

J. CHEM. INF. COMPUT. SCI., vol. 23, 1983, pages 109-117, American Chemical Society, Washington, D.C., US; Y. KUDO et al.: "Chemical substance retrieval system for searching generic representations. 1. A prototype system for the gazetted list of existing chemical substances of Japan"

Description

TECHNICAL FIELD

This invention relates to a method of storing chemical structure data in a storage device and searching said chemical structure data using a query chemical structure by examining the match or analogy between the said query structure with the structure data stored.

BACKGROUND OF THE ART

Recently various information including patent information is more and more handled by computer. Textual data, which consist of alphabets and numerals, such as patent claim information or technical information are now stored as a database and specific pieces of information are easily retrieved by searching the database. In a textual database, keywords are picked up from each piece of information, generally called a record, and those keywords are sorted alphabetically in the database as an inverted file. A search is conducted by combining the record list of each keyword using the Boolean operation with AND, OR, or NOT logical operators. The basic idea of this method was introduced early in the 1960's. The first computer system using this method was introduced in the 1970's in the United States. Most of the current online information retrieval systems use this type of textual data retrieval method.

On the other hand, storage and retrieval of chemical structure information, which is a graphic data in nature, was not so easy to achieve as that of textual data. Handling of chemical susbstance data is discussed in a book, "Chemical Information System", edited by J. E. Ash and E. Hyde, Ellis Horwood Ltd., 1975. A related U. S. patent is 4,085,443 by Dubois et al. It was only early in the 1980's when chemical structure storage and retrieval systems were available commercially. An inverted file which is used to handle textual information is not applicable to graphic data such as chemical structure data. Rather it is necessary to compare atoms and bonds of a query chemical structure with those of each chemical structure stored in a database to find a match between those structures. In order to do this comparison, it is necessary to create and keep so-called connection tables for all query and file structures. Since this comparison requires tracking atom connections one by one, it is usually called an iterative search. The iterative search consumes much computer time and affects overall search time considerably. It is a necessity to minimize the number of candidate file structures to which iterative searches are to be conducted by screening out most "unwanted" structures. The screening is achieved by checking the presence or absence of particular chemical characteristics called screens requested by the query structure. For example, if the query structure contains a nitrogen atom, any file structures which do not have nitrogens will be screened out. In a current commercial system, screens are created automatically by a computer, when a query structure is created through an interactive session on a remote graphic terminal.

Thus, the current chemical structure search systems can handle specifically defined structures which are usually found in technical journals. On the other hand, generic expression of chemical structures is widely used in patent claims to widen the coverage of those claims. Specific examples of the generic expression are:

Alkyl groups with C1-C5 chain.

Aromatic rings (i.e., benzene or naphthalene)

Heterocyclic groups (i. e., rings containing one or more non-carbon atoms) with ring size of 5 or 6.

The generic expression often covers thousands or millions of specific chemical structures, and allows expansion of the scope of a claim without identifying each structure specifically. Since chemical substances themselves are patentable in most countries, it is very important to store and search the generic chemical structures. Current status of handling of generic chemical structures is discussed thoroughly in the following references.

(1) "Computer Storage and Retrieval of Generic Chemical Structures in Patents. 1. Introduction and General Strategy" by M. F. Lynch, S. M. Welford, and J. M. Barnard, J. Chem. Inf. Comput. Sci., 1981, (21), 148-150.

(2) "Computer Storage and Retrieval of Generic Chemical Structures in Patents. 2. GENSAL, a Formal language for the Description of Generic Chemical Structures" by J. M. Barnard, M. F. Lynch, and S. M. Welford, J. Chem. Inf. Comput. Sci., 1981, (21), 151-161.

(3) "Computer Storage and Retrieval of Generic Chemical Structures in Patents. 3. Chemical Grammars and their Role in the Manipulation of Chemical Structures" by S. M. Welford, M. F. Lynch, and J. M. Barnard, J. Chem. Inf. Comput. Sci., 1981, (21), 157-163.

(4) "Computer Storage and Retrieval of Generic Chemical Structures in Patents. 4. An Extended

Connection Table Representation for Generic Structures" by J. M. Barnard, M. F. Lynch, and S. M. Welford, J. Chem. Inf. Comput. Sci., 1982, (22), 160-164.

(5) "Chemical Substance Retrieval System for Searching Generic Representations. 1. A Prototype System for the Gazetted List of Existing Chemical Substances of Japan" by Y. Kudo and H. Chihara, J. Chem. Inf. Comput. Sci., 1983, (23), 109-117.

(6) "Computer Storage and Retrieval of Generic Chemical Structures in Patents. 5. Algorithmic Generation of Fragment Descriptors for Generic Structure Screening" by S. M. Welford, M. F. Lynch, and J. M. Barnard, J. Chem. Inf. Comput. Sci., 1984 (24), 57-66.

(7) "Computer Storage and Retrieval of Generic Chemical Structures in Patents. 6. An Interpreter Program for the Generic Structure Description Language GENSAL" by J. M. Barnard, M. F. Lynch, and S. M. Welford, J. Chem. Inf. Comput. Sci., 1984 (24), 66-71.

(8) "A Relaxation Algorithm for Generic Chemical Structure Screening" by A. Von Scholley, J. Chem. Inf. Comput. Sci., 1984 (24) 235-241.

Because of their complexity, no system could handle generic chemical structures successfully until now, except that three approaches were made to solve the problem partially.

(APPROACH-A)

One approach is to store specific structures expressed by the generic structures. Practically, a database containing structure information of substances specifically identified in patent examples is widely used. One example is the Registry File of CAS ONLINE. But patent examples usually describe only a portion of the generic structures in a claim, and thus it is not usually true that the combination of all chemical structures in the examples corresponds to the claimed generic expression. It is certainly not practical to expand generic structures into component specific structures, since the number of specific structures derived from one generic structure easily explodes to millions. For example, an expression C4-C5 alkyl group represents 12 specific alkyl radicals. If a generic structure contains three of this expression, the combination results in 12 x 12 x 12, or 1728 specific structures.

(APPROACH-B)

The other approach is to define codes for various chemically significant units, such as rings, chains and functional groups, and search the structure via those codes, like keywords of textual databases. Examples are the World Patent index of Derwent or Comprehensive Database of IFI. In this approach, the expression C4-C5 alkyl group may be coded into two keywords, C4 and C5. Thus even very complex generic structures can be coded fairly simply. One shortcoming of this approach is that a searcher has to know the coding rule and use the necessary codes explicitly. For example, in searching for a propyl group, one has to specify keywords both PROPYL and C3 ALKYL. But a bigger problem is that the coding system cannot express the connection between the chemical units successfully. This results in large numbers of irrelevant answers, which are usually called noise. Often more than 90% of the answer structures are noise. Another disadvantage is that since the file has no connection tables, or exact representation of chemical structures, it is unable to search by structures, as one can do in the system based on specific connection tables. Thus the searcher needs to learn how to use the code system to code a query structure effectively. Apparently, this prevents the system from wide use.

(APPROACH-C)

A latest approach, which is described in the reference (5) by Kudo and Chihara, is trying to solve the problem by defining a connection table made solely of generic nodes such as heterocyclic or alkyl groups.

In this reference, the authors define three representations, i.e., the Q, R, and S representations. Here the S representation is the connection table made of specific atoms as defined by the regular atomic table, and the Q representation is the connection table made of generic atom groups such as ZZ for a heterocyclic ring. Thus the Q representation is a derivative of the S representation. The R representation is the attribute data for Q, e.g. the number of heteroatoms for ZZ or a heterocyclic ring. The method of searching here is to compare each level of representations separately. When a query is input, each level of representations is generated automatically (page 110, column 2, lines 51-55). The query structure and stored structures are compared at individual levels of representations, i.e. S versus S, and Q versus Q (page 114, column 2, lines 25-31). There was no intention to compare representation of different levels.

DISCLOSURE OF THE INVENTION

None of the three current approaches described above fulfills the needs of generic structure searching. In addition, in the true generic structure search system, a query structure itself may have many generic expressions as well as the structures stored in a database. This type of search, namely generic structure searching using a generic structure query, has been considered almost impossible.

Among many problems in developing a generic structure search system based on chemical structures rather than on code systems, the biggest one is how to represent such generic structure units as alkyl, heterocyclic, etc., usually described as textual information in the patent disclosure, in a structure connection table as searchable information. In comparing a query structure with a file structure, it is necessary to find a match between nonidentical expressions. For example, we find a match when the query requires the presence of C1-C5 alkyl and a file structure has a C4-C7 alkyl, because the components of C4 and C5 of the file structure satisfy the requirement of the query. Also there is a match between a query specifying a nitrogen-containing ring and a file structure containing a six-membered (ring size of six) heterocycle. The claimed invention is intended to solve the above problems in developing the generic structure search system and provide it for practical use.

Accordingly, there is provided a method of storing chemical structure data in a storage device and searching said chemical structure data using a query chemical structure by examining the match or analogy between the said query structure and the structure data stored, wherein said method comprises the steps of:

assigning numbers to each chemical unit, which can be either an atom of a specific or generic element, or a generic group of atoms, of each structure to be stored, storing the numbers of chemical units to which the said chemical unit is chemically connected in a connection register, storing the attribute data, which describe the chemical characteristics of the said chemical unit, in an attribute register,

then assigning numbers to each chemical unit, which can be either an atom of a specific or generic element, or a generic group of atoms, of the query chemical structure to be used to search the stored chemical structure data, storing the numbers of chemical units to which the said chemical unit is chemically connected in a connection register, storing the attribute data, which describe the chemical characteristics of the said query chemical unit, in an attribute register,

then examining the match or analogy of the query structure and each stored chemical structure by comparing chemical units of the query chemical structure with the chemical units of the stored chemical structure by matching the attributes of the chemical unit of the stored chemical structure and the attributes of the corresponding chemical units of the query structure according to a mathematical condition defined in advance, either with or without matching the specific or generic element type of the chemical unit of the stored chemical structure with the specific or generic element type of the chemical unit of the query chemical structure,

wherein said attribute is made of a predefined vector, each column of which represents specified chemical characteristics of chemical structures, and wherein the existence or absence of the specified chemical characteristics is expressed by the existence or absence of a predefined value in the column which specified the said chemical characteristics, or alternatively is expressed by the absence or existence of a predefined value in the column which specified the said chemical characteristics, and

wherein said mathematical condition is the result of a series of logical operations conducted between a vector or vectors of a query chemical structure and a vector or vectors of a stored chemical structure.

BRIEF DESCRIPTION OF THE DRAWINGS

Fig. 1 illustrates a search algorithm based on the connectivity stack method;

Fig. 2 illustrates a possible process of examining a match of a query node with the corresponding file node;

Fig. 3 illustrates a possible process of examining matches of connections around a query node and those around the corresponding file node;

Fig. 4 illustrates a possible process of examining matches of attributes between a query node and the corresponding file node; and

Fig. 5 shows an apparatus which would operate in accordance with our invention.

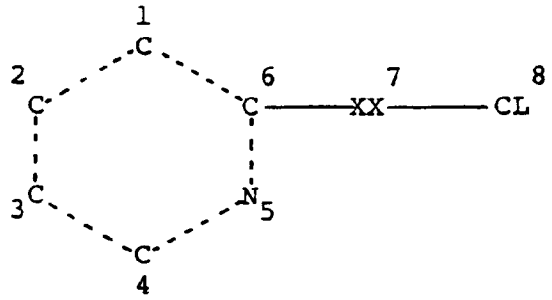
DESCRIPTION OF BEST MODE FOR CARRYING OUT THE INVENTION

The invention will now be described in detail in conjunction with a preferred mode thereof.

(1) SAMPLE GENERIC CHEMICAL STRUCTURES

The stored structure (Structure 1) is represented by the connection table (Table 1) and the attribute table (Table 2), and the query structure (Structure 2) is represented by the connection table (Table 3) and the attribute table (Table 4).

STRUCTURE 1.



(XX is an oxygen-containing ring system consisting of one or two component rings with the ring size of five or six.)

The search process is described step by step in Table 5 and the chain attribute is described in Table 6.

First all the atoms (including generic atoms and excluding hydrogens) which will be called nodes hereafter, are numbered at will. In this example, these nodes correspond to the chemical unit described earlier.

TABLE 1.

NO.	ELEMENT	ATTRIBUTE	Neighbor NODE NO.	BOND	Neighbor NODE NO.	BOND	Neighbor NODE NO.	BOND	Neighbor NODE NO.	BOND
1	C	6-membered ring with nitrogen	2	4	6	4				
2	C	6 membered ring with nitrogen	1	4	3	4				
3	C	6-membered ring with nitrogen	2	4	4	4				
4	C	6-membered ring with nitrogen	3	4	5	4				
5	N	6-membered ring with nitrogen	4	4	6	4				
6	C	6-membered ring with nitrogen	1	4	5	4	7	1		
7	XX	5- or 6-membered ring with oxygen	6	1	8	1				
8	CL	halogen	7	1						

A table is created with each row representing each node of the structure and the first column representing the node number. The second column shows the element type of the node, the third column shows the attribute code of the node, the fourth column shows the number of a node to which the current node is connected, and the fifth column shows the bond value of the connection above. The bond value is coded such that a single bond is represented by 1, double bond, 2, triple bond, 3, aromatic bond, 4. In Structure 1, a single bond is represented by a straight line and an aromatic bond by a broken line. It is important that columns 4 and 5 form a pair to describe connection information. Similarly, the pairs of columns 6 and 7, 8 and 9, and 10 and 11 each describe connections of the current node to the other nodes. Although there are only four pairs of connection information in Tables 1 and 3, one may add more columns if more than four connections are expected.

Although attribute data are expressed by texts in Tables 1 and 3 to enhance comprehension, it is usually represented by bit-maps as in Tables 2 and 4.

TABLE 2.

NO.	RING ATOM				RING SIZE				RING NUMBER MIN-IMUM			MAX-IMUM		
	O	S	N	A	4	5	6	>6	1	2	3	1	2	3
1	0	0	1	0	0	0	1	0	1	0	0	1	1	1
2	0	0	1	0	0	0	1	0	1	0	0	1	1	1
3	0	0	1	0	0	0	1	0	1	0	0	1	1	1
4	0	0	1	0	0	0	1	0	1	0	0	1	1	1
5	0	0	1	0	0	0	1	0	1	0	0	1	1	1
6	0	0	1	0	0	0	1	0	1	0	0	1	1	1
7	1	0	0	0	0	1	1	0	1	1	0	1	1	1

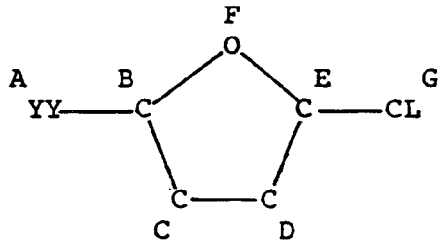
Thus pre-defined chemical characteristics of each chemical unit, or node in this example, are assigned a position in the bit-map vector, which is widely used in computer programming. If a particular chemical characteristic exists in the node, the position of the bit-map is filled by number 1, otherwise it is filled by number 0. Although there is no limitation in the length of the bit-map, or the number of positions in it, the range may be between 4 to 4096, preferably 8 to 512, to make best use of computer resources. These bit-maps are easily handled by ordinary electronic computer systems. Thus the generic node expression in the node 7 of Table 1 is expressed as follows. First a dummy element code XX is put in the row 7, column 2 of Table 1, showing that the node 7 is a generic node. Then each column (bit-map position) of row 7 of Table 2 is filled with either 1 or 0 depending on the presence or the absence of the corresponding chemical characteristics. For such numeric characteristics as the count of carbon atoms or the number of rings, two sets of columns are usually prepared to represent both the maximum and minimum values. In this context, a minimum value simply means the actual value may not be lower than it, and a maximum value means the actual value may not be larger than it. Thus in Table 2, when the number of rings is one, the column of minimum value of 1 is filled with 1, and the columns of maximum value of 1, 2 and 3 are filled with 1. When the number of rings is in the range of 1 to 2 as the node 7, the columns corresponding to the minimum value of 1 and 2, and the maximum value of 1, 2 and 3 are filled with 1. The latter case may be interpreted by combining the bit-maps corresponding to the number of rings, 1 and 2, by the logical operator OR.

Since only the ring-related attributes are shown in Table 2 to make the explanation simple, the node 8 does not have any attribute here. The important point is that this attribute bit-map is generated for all nodes of the structure regardless of whether they are generic or not. It is convenient to define separate attributes for rings and chains. Although it is very easy to distinguish if a generic node is a ring or chain in most cases, it may be necessary to divide a structure into two, namely the one with the ring node and the one with a chain node when the distinction is not very clear.

The search is conducted by comparing one by one the nodes of the query structure (Structure 2) with

the nodes of the stored structure (Structure 1).

STRUCTURE 2.



(YY is a nitrogen-containing heterocycle. The number of component rings is unlimited.)

TABLE 3.

NO.	ELEMENT	ATTRIBUTE	Neighbor NODE NO.	BOND	Neighbor NODE NO.	BOND	Neighbor NODE NO.	BOND	Neighbor NODE NO.	BOND
A	YY	heterocycle with nitrogen	B	1						
B	C	5-membered ring with oxygen	A	1	C	1	F	1		
C	C	5-membered ring with oxygen	B	1	D	1				
D	C	5-membered ring with oxygen	C	1	E	1				
E	C	5-membered ring with oxygen	D	1	F	1	G	1		
F	O	5-membered ring with oxygen	B	1	E	1				
G	CL	halogen	E	1						

The nodes of the query structure are assigned alphabet values rather than numerals. The generic node A of the query structure is assigned a dummy element value of YY. In ordinary chemical structure search systems, it is convenient to examine the match of element type first and then the matches of connection, which consists of the node number of the connecting atom and the bond value. In this invention the matches of attributes are most important. Although the examination steps described in Table 5 show that the element matches precede the attribute match, this order is not critical. It is also possible to eliminate the element matches altogether.

The attributes of the query structure are shown in Table 4. Since the number of rings of the node A is undefined (more than one), only the column of the minimum value of 1 is assigned the number 1. Also no attribute is assigned for the size of the ring.

TABLE 4.

NO.	RING ATOM				RING SIZE				RING NUMBER MIN-IMUM			MAX-IMUM		
	O	S	N	A	4	5	6	>6	1	2	3	1	2	3
A	0	0	1	0	0	0	0	0	1	0	0	0	0	0
B	1	0	0	0	0	1	0	0	1	0	0	1	1	1
C	1	0	0	0	0	1	0	0	1	0	0	1	1	1
D	1	0	0	0	0	1	0	0	1	0	0	1	1	1
E	1	0	0	0	0	1	0	0	1	0	0	1	1	1
F	1	0	0	0	0	1	0	0	1	0	0	1	1	1

A match is determined when every column of the attribute bit-map of a stored structure node fulfills the corresponding column of the bit-map of a query structure node, i. e., if a column of the query node is one, then the corresponding column of the stored structure node MUST be one. On the other hand, if the column is zero, the corresponding column of the stored structure node can be either one or zero. It is possible to exchange the value one and zero, depending on the system architechture. It is also possible to define values other than 1 and 0.

When one uses a standard bit-map with values 1 and 0, it is very easy to compare the bit-map of the query structure with the bit-map of the stored structure using a series of logical operations usually available on an electronic computer. An example step is to NOT the bit-map of the stored structure, and then to AND the result with the bit-map of the query structure. If the result of the AND operation is a zero vector, or a bit-map filled with zero, then the positions of the bit-map of the stored structure, corresponding to the positions of the bit-map of the query structure where their value is 1, should be 1. For example, the result of the NOT operation of the bit-map row 1 of Table 2 (0010 0010 100 111) is 1101 1101 011 000. When it is ANDed with the bit-map row A of Table 4 (0010 0000 100 000), the result is 0000 0000 000 000. Thus all the chemical characteristics defined on the bit-map row A (of Table 4) by bits 1 do exist in the bit-map row 1 (of Table 2). On the other hand, the result of an AND operation with the bit-map row B (1000 0100 100 111) is 1000 0100 000 000 and is not all-zero. Thus we know that the ring atom 0 and ring size 4 do not exsist in the bit-map row 1.

Another alternative may be to compare the attribute columns to see if one is larger than the other. The invention described here is characteristic in that the comparison is made using some mathematical methods discussed above rather than by a simple code or keyword (text) matching.

(2) SEARCH PROCESS

We chose here substructure searching, where a match is successful when a given query structure is logically involved or embedded in a candidate structure. Thus an examination is over either when all query nodes are matched with corresponding file nodes, or when it is found that such a query-file node correspondence does not exist.

All the query nodes should be examined, but not all the file nodes, as long as the above condition is fulfilled.

One may choose other searches than substructure searching. Several differences exist between different types of search in the process described above. But the use of the attribute is valid in any process.

An apparatus which would operate in accordance with our search process is illustrated in Fig. 5. First the query structure is input by the graphic structure input device 1, and converted into a connection table. The connection table data are set to each component register row of the connection table register 8. Usually there is more than one structure in a file, and thus the search controller 3 reads a connection table

and set data to each component register row of the connection table register 9. The count of the file connection table is incremented whenever a new connection table is read. Comparison of a query structure and a file structure is controlled by the node comparison status controller register 5.

Several method are described elsewhere to match the connecting nodes. Although we chose the connectivity stack method here, it is certainly not the sole method of achieving the goal.

The specific search process conducted by the node comparison status controller based on the connectivity stack method is illustrated in Figure 1.

Here nodes of the query structure (hereafter called query nodes) are numbered by i. By definition, $0<i<=m$, where m is the count of the query nodes. Similarly, nodes of a candidate structure in the file (hereafter called file nodes) are numbered by j. Also by definition, $0<j<=n$, where n is the count of the file nodes. A vector P(i) is defined to store the number of a file node which matches the query node i. Another vector M(j) is defined to mark if the file node j is already matched with a query node. If it is, its value is 1. Otherwise it is 0. Since a generic file node may be matched with multiple query nodes, the M(j) value of a generic file node is kept 0. An examination of a file structure against a query structure is conducted by comparing every query node ($0<i<=m$) with every file node ($0<j<=n$), except those file nodes already matched with query nodes, for which M(j) equals 1.

The match of a query node and a file node is examined by comparing their attributes read to the attribute comparison register 24. elements where no generic nodes are involved read to the element comparison register 25, connections with nodes already matched read to the connection comparison register 26, and other conditions, if necessary. It should be noted that by using only attributes a match can be examined between a generic node and a specific (single) node. The process is illustrated in Figure 2.

The connections of a query node and those of a file node with those nodes already matched may be compared as illustrated in Figure 3 by reading the connection data from the component register 10 to 16 of the query connection table register 8 for a particular query node, and the component registers 17 to 23 of the file connection table register 9 for a particular file node. If the query node A has a connection with a query node B already matched with a file node Y, then the file node X, which is now compared with the query node A, must have a connection with the file node Y. In addition, the bond type between X and Y must be the same with that of A and B. If the file node X is generic, i.e., can be matched with multiple query nodes, there is a chance both A and B matches X. Then the connection between A and B need not be examined as long as both A and B belong to the same ring or chain. If a connecting query node B is generic, a connecting file node need not be necessarily Y, which was matched with B. If a connecting file node of X, which is Z, is found to be in the same ring or chain, it is determined that the connection is matched.

Attributes of a query structure and a stored structure mentioned in Figure 2 may be compared by the step as illustrated in Figure 4. A series of logical operations is applied as described earlier.

Although the steps described using the Figures 1, 2, 3, and 4 represent a typical search procedure of this invention, certainly it is not the sole method to achieve the goal. Any procedure which uses the chemical structure attributes defined in this invention is a valid alternative.

(3) STEP-BY-STEP ILLUSTRATION

A step-by-step illustration is given in Table 5, using the example Structures 1 and 2, and Tables 1, 2, 3 and 4. As described in Table 5, this method develops a vector with the node numbers of the query structure as components, and assigns the nodes of the stored structure one by one, comparing the connection with other nodes.

TABLE 5.

STEP	MODE OF QUERY	A	B	C	D	E	F	G	CON-NECT	MATCH	ELE-MENT	ATTRI-BUTE	BOND
	1	1							1-A	○	*	○	
	2	1	2						2-B	×	○	×	
	.												
	.												
	.												
	8	1	8						8-B	×	×		
	9	2							2-A	○	*	○	
	10	2	1						1-B	×	○	×	
	.												
	.												
	.												
	39	6							6-A	○	*	○	
	40	6	1						1-B	×	○	×	
	.												
	.												
	.												
	46	6	7						7-B	○	**	○	○
	47	6	7	1					1-C	X	0	X	
	.												
	52	6	7	7					7-C	0	**	0	0***
	.												
	72	6	7	7	7	7	7	7	7-G	×	**	×	.
	73	6	7	7	7	7	7	8	8-G	○	○	○	○

* Element is not examined for the generic node A
** Element is not examined for the generic node 7
*** Presence of connection assumed between nodes belonging to a same generic node

First the node 1 of the stored structure matches with the node A of the query structure by their attributes. The element value is not examined since the node A is generic. No connection is examined for the first node. Then by comparing the node 2 and the node B, one finds that their attributes do not match (the node B has a 5-membered ring while the node 2 has a 6-membered ring), although the element is the same. By exchanging the node 2 by the rest of the nodes of the stored structure, no matches are found at last at the step 8. Then one concludes that the initial assignment of the node 1 and the node A is irrelevant. Thus one begins comparing the node 2 with the node A instead. A successful match is found when the node 6 is assigned to the node A, and the node 7 is compared with the node B. Since the node 7 is generic, only the attribute is examined. The attribute of the node 7 tells that it has a 5- or 6-membered ring with oxygen atoms, which fulfills the condition of the node B. Since the node B is connected with the node A, it is required that the node 7 be connected to the corresponding node 6. This requirement is also fulfilled fortunately. Since the bond values of those two connections are the same, the single bond, the match of the node 7 and B are confirmed.

Then the node C, D, and E of the query structure all match with the node 7. The match is confirmed by the fact that node C, D, and E are in the same ring system, because the generic node 7 must be a single ring system. Duplicate matches are possible, since the generic node 7 is considered to consist of multiple nodes. Finally, since the node 8 matches with the node G in every respect, it is concluded that the stored structure 1 matches the query structure 2.

(4) EXPLANATION OF THE EXAMPLE

As is described above, by using the expanded connection tables and the attribute tables, the matches of specifically defined nodes and generic nodes are easily examined, and thus the matches of the query and stored structures containing both specific and generic nodes are determined. Ordinary electronic computers are easily programmed to conduct the search.

In the above example the comparison is made under the requirement that a query structure is either identical with or embedded in the stored structure. This type of search is called sub-structure searching and is widely used today on specific (not generic) substance databases. But the application of this invention is not limited to sub-structure searching but to the exact matching of structures or to such searches where the requirement is that a stored structure should be embedded in the query structure.

Although only the ring attribute was described in detail in the above example, the chain attribute is also important. Chains are often expressed generically as C3-C5. When a node of a stored structure is C3-C5 and a node of a query structure is C4-C7, attributes of each node are expressed as in Table 6.

TABLE 6.

	CHAIN LENGTH ATTRIBUTE															
	MINIMUM								MAXIMUM							
	1	2	3	4	5	6	7	8	1	2	3	4	5	6	7	8
C3-C5 in stored structure	1	1	1	1	1	0	0	0	0	0	1	1	1	1	1	1
C4-C7 in query structure	1	1	1	1	0	0	0	0	0	0	0	0	0	0	1	1

For the stored structure node, the attribute bit-map is the result of a logical OR operation of the attribute of C3, C4, and C5. On the other hand the attribute of the query node is the result of an AND combination of the attribute of C4, C5, C6, and C7. In this example, the node of the stored structure fulfills the attribute requirement of the query node, and the match is found.

EFFECT OF THE INVENTION

Since this invention uses the extended connection table and the attribute table, the representation is compact. It requires much less computer storage than the method using whole set of structures derived from a generic structure as described in the method A. Since this invention is based on traditional connection tables, the search system can be developed based on structure graphics. Thus a searcher can build a query chemical structure image familiar to chemists on a graphic terminal and conduct a search using it. There is no need to learn complex coding rules as described in the method B.

This invention breaks a generic expression down to chemically-significant units and represents them by a finite number of attribute positions on a bit-map. This allows one to find a match between non-identical expressions such as a C3-C5 alkyl, and a C4-C7 alkyl, or a nitrogen-containing ring and a six-membered heterocycle. Since these attributes are generated for both specifically-defined nodes and generic nodes, the comparison between those two types of nodes is easy.

Since it is usually easy to generate the attributes algorithmically from the normal generic expression as well as the specific expression, the method of the invention is easily implemented on a computer. It is important, too, that the method of the invention is compatible with the currently available specific structure search system using connection tables so that mixed handling of both generic structure and specific structure data is possible.

As a conclusion, this invention is particularlly useful in storing and retrieving complex chemical structures using a computer.

INDUSTRIAL UTILITY OF THE INVENTION

This invention is intended to be used in information storage and retrieval systems, specifically in chemical and chemical structure information storage and retrieval systems. This invention is specifically intended to provide an effective way to store and search complex chemical structure data, especially those

chemical structure expressions widely known as "generic" chemical structures, on a computer.

Claims

1. A method of storing chemical structure data in a storage device and searching said chemical structure data using a query chemical structure by examining the match or analogy between the said query structure and the structure data stored, which comprises the steps of:

   assigning numbers to each chemical unit, which can be either an atom of a specific or generic element, or a generic group of atoms, of each structure to be stored, storing the numbers of chemical units to which the said chemical unit is chemically connected in a connection register, storing the attribute data, which describe the chemical characteristics of the said chemical unit, in an attribute register,

   then assigning numbers to each chemical unit, which can be either an atom of a specific or generic element, or a generic group of atoms, of the query chemical structure to be used to search the stored chemical structure data, storing the numbers of chemical units to which the said chemical unit is chemically connected in a connection register, storing the attribute data, which describe the chemical characteristics of the said query chemical unit, in an attribute register,

   then examining the match or analogy of the query structure and each stored chemical structure by comparing chemical units of the query chemical structure with the chemical units of the stored chemical structure by matching the attributes of the chemical unit of the stored chemical structure and the attributes of the corresponding chemical units of the query structure according to a mathematical condition defined in advance, either with or without matching the specific or generic element type of the chemical unit of the stored chemical structure with the specific or generic element type of the chemical unit of the query chemical structure,

   wherein said attribute is made of a predefined vector, each column of which represents specified chemical characteristics of chemical structures, and wherein the existence or absence of the specified chemical characteristics is expressed by the existence or absence of a predefined value in the column which specified the said chemical characteristics, or alternatively is expressed by the absence or existence of a predefined value in the column which specified the said chemical characteristics, and

   wherein said mathematical condition is the result of a series of logical operations conducted between a vector or vectors of a query chemical structure and a vector or vectors of a stored chemical structure.

2. The method claimed in claim 1 wherein said attribute vectors consist at least of a vector or vectors describing ring characteristics of chemical structures and a vector or vectors describing chain characteristics of chemical structures.

3. The method claimed in claim 2 wherein said logical operations are conducted using electronic computers or electronic data processors.

Revendications

1. Un procédé d'enregistrement de données de structure chimique dans une mémoire et de recherche desdites données de structure chimique en utilisant une structure chimique de recherche par examen de la concordance ou analogie entre ladite structure de recherche et les données de structure enregistrées, procédé qui comprend les étapes consistant à :
   - affecter des nombres à chaque unité chimique, qui peut être soit un atome d'un élément spécifique ou générique, soit un groupe générique d'atomes, de chaque structure à enregistrer, enregistrer les nombres d'unités chimiques auxquelles ladite unité chimique est chimiquement liée dans un registre de liaison, enregistrer les données d'attributs, qui définissent les caractéristiques chimiques de ladite unité chimique, dans un registre d'attributs,
   - ensuite affecter des nombres à chaque unité chimique, qui peut être soit un atome d'un élément spécifique ou générique, soit un groupe générique d'atomes, de la structure chimique de recherche à utiliser pour rechercher les données de structure chimique enregistrées, enregistrer les nombres d'unités chimiques auxquelles ladite unité chimique est liée chimiquement dans un registre de liaison, enregistrer les données d'attributs, qui définissent les caractéristiques chimiques de ladite unité chimique de recherche, dans un registre d'attributs,
   - ensuite examiner la concordance ou analogie de la structure de recherche et de chaque structure

chimique enregistrée en comparant les unités chimiques de la structure chimique de recherche avec les unités chimiques de la structure chimique enregistrée par mise en correspondance des attributs de l'unité chimique de la structure chimique enregistrée et des attributs des unités chimiques correspondantes en concordance avec une condition mathématique définie à l'avance, soit avec soit sans mise en correspondance du type d'élément spécifique ou générique de l'unité chimique de la structure chimique enregistrée avec le type d'élément spécifique ou générique de l'unité chimique de la structure chimique de recherche,

- procédé où ledit attribut est constitué d'un vecteur prédéfini, dont chaque colonne représente des caractéristiques chimiques spécifiées de structures chimiques, et où l'existence ou l'absence des caractéristiques chimiques spécifiées est exprimée par l'existence ou l'absence d'une valeur prédéfinie dans la colonne qui spécifiait lesdites caractéristiques chimiques, ou bien en variante elle est exprimée par l'absence ou l'existence d'une valeur prédéfinie dans la colonne qui spécifiait lesdites caractéristiques chimiques, et
- où ladite condition mathématique est le résultat d'une série d'opérations logiques effectuées entre un vecteur ou des vecteurs d'une structure chimique de recherche et un vecteur ou des vecteurs d'une structure chimique enregistrée.

2. Le procédé tel que revendiqué dans la revendication 1, dans lequel lesdits vecteurs d'attributs se composent d'au moins un ou plusieurs vecteurs décrivant des caractéristiques de noyaux de structures chimiques, et d'un ou plusieurs vecteurs décrivant des caractéristiques de chaînes de structures chimiques.

3. Le procédé tel que revendiqué dans la revendication 2, dans lequel lesdites opérations logiques sont effectuées en utilisant des ordinateurs électroniques ou des processeurs électroniques de données.

Patentansprüche

1. Verfahren zum Speichern chemischer Strukturdaten in einem Speichergerät und zum Suchen der chemischen Strukturdaten unter Verwendung einer chemischen Suchstruktur durch Überprüfen der Übereinstimmung oder Analogie zwischen der Suchstruktur und den gespeicherten Strukturdaten, enthaltend folgende Verfahrensschritte:
jeder chemischen Einheit, die entweder ein Atom eines speziellen oder generischen Elementes oder eine generische Gruppe von Atomen sein kann, jeder zu speichernden Struktur werden Nummern zugeordnet, die Nummern der chemischen Einheiten, mit denen die chemische Einheit chemisch verbunden ist, werden in einem Verbindungsregister gespeichert, und die Attributdaten, die die chemischen Charakteristika der chemischen Einheit beschreiben, werden in einem Attribut-Register gespeichert,
dann werden jeder chemischen Einheit, die entweder ein Atom eines speziellen oder generischen Elementes oder eine generische Gruppe von Atomen sein kann, der zum Suchen der gespeicherten chemischen Strukturdaten zu verwendenden chemischen Suchstruktur Nummern zugeordnet, die Nummern der chemischen Einheiten, mit denen die chemische Einheit chemisch verbunden ist, werden in einem Verbindungsregister abgespeichert und die Attributdaten, die die chemischen Eigenschaften der chemischen Sucheinheit beschreiben, werden in einem Attribut-Register abgespeichert,
dann wird die Übereinstimmung oder Analogie der Suchstruktur und jeder gespeicherten chemischen Struktur dadurch überprüft, daß chemische Einheiten der chemischen Suchstruktur mit den chemischen Einheiten der abgespeicherten chemischen Struktur verglichen werden, indem die Attribute der chemischen Einheit der gespeicherten chemischen Struktur und die Attribute der entsprechenden chemischen Einheiten der Suchstruktur entsprechend einer vorher bestimmten mathematischen Bedingung verglichen werden, entweder mit oder ohne Übereinstimmung der speziellen oder generischen Elementart der chemischen Einheit der gespeicherten chemischen Struktur mit der speziellen oder generischen Elementart der chemischen Einheit der chemischen Suchstruktur,
wobei das Attribut aus einem vordefinierten Vektor besteht, von dem jede Spalte spezifizierte chemische Eigenschaften chemischer Strukturen darstellt, und wobei das Vorhandensein oder Fehlen der spezifizierten chemischen Eigenschaften durch das Vorhandensein oder Fehlen eines vorbestimmten Wertes in der Spalte ausgedrückt wird, die die chemische Charakteristika spezifizierte, oder aber durch das Fehlen oder Vorhandensein eines vorbestimmten Wertes in der Spalte ausgedrückt wird, die die chemische Charakteristika spezifizierte, und
wobei die mathematische Bedingung das Ergebnis einer Reihe von logischen Operationen ist, die

zwischen einem Vektor oder Vektoren einer chemischen Suchstruktur und einem Vektor oder Vektoren einer gespeicherten chemischen Struktur durchgeführt werden.

2. Verfahren nach Anspruch 1, bei dem die Attribut-Vektoren mindestens aus einem Vektor oder Vektoren, die Ringeigenschaften chemischer Strukturen beschreiben, und einem Vektor oder Vektoren bestehen, die Kettencharakteristika chemischer Strukturen beschreiben.

3. Verfahren nach Anspruch 2, bei dem die logischen Operationen unter Verwendung elektronischer Computer oder elektronischer Datenprozessoren durchgeführt werden.

FIG. 1

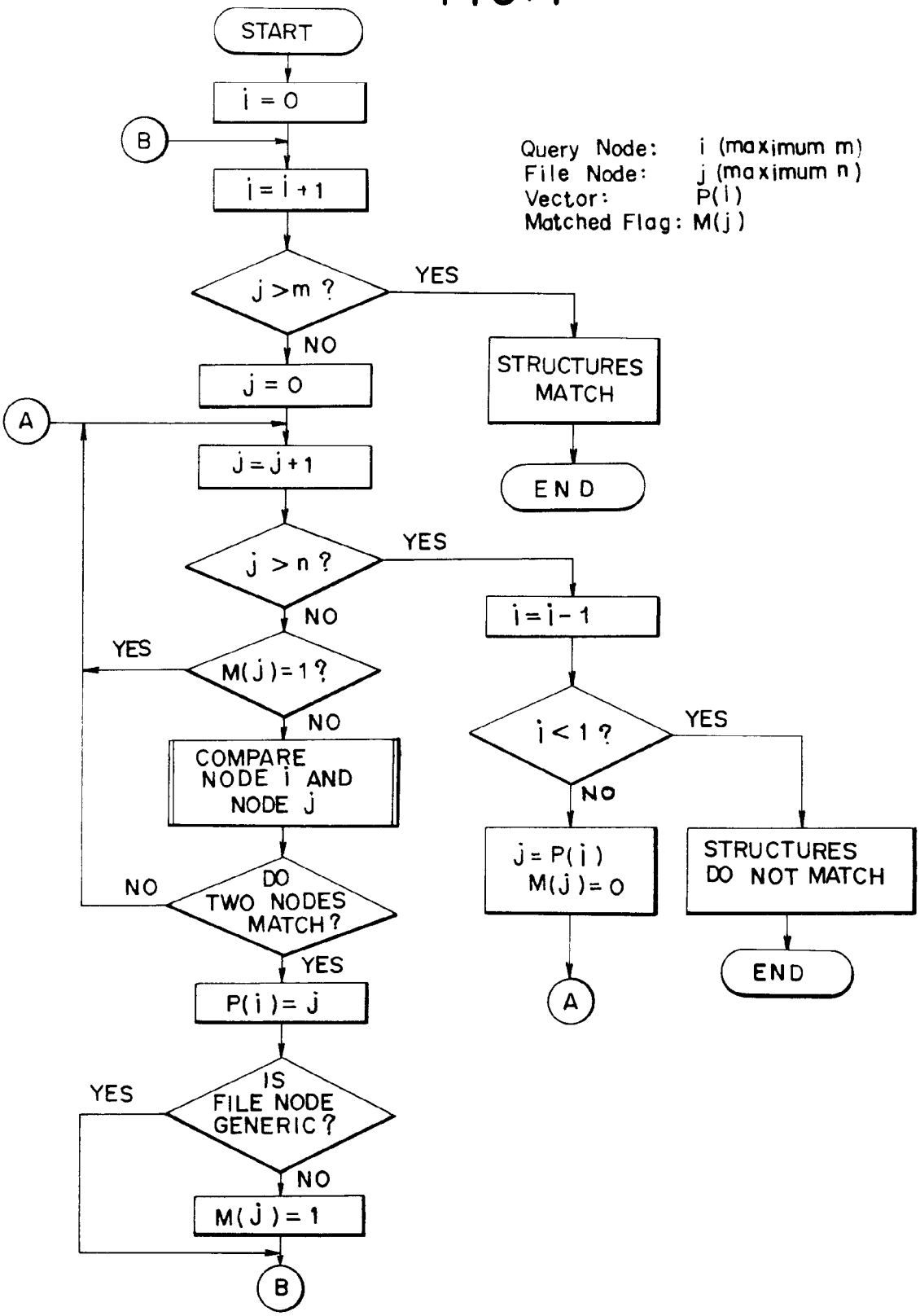

START
i = 0
B
i = i + 1
j > m ?
YES
NO
STRUCTURES MATCH
END
j = 0
A
j = j + 1
j > n ?
YES
NO
i = i - 1
M(j) = 1 ?
YES
NO
i < 1 ?
YES
NO
COMPARE NODE i AND NODE j
j = P(i)
M(j) = 0
STRUCTURES DO NOT MATCH
END
DO TWO NODES MATCH?
NO
YES
A
P(i) = j
IS FILE NODE GENERIC?
YES
NO
M(j) = 1
B
Query Node: i (maximum m)
File Node: j (maximum n)
Vector: P(i)
Matched Flag: M(j)

FIG. 2

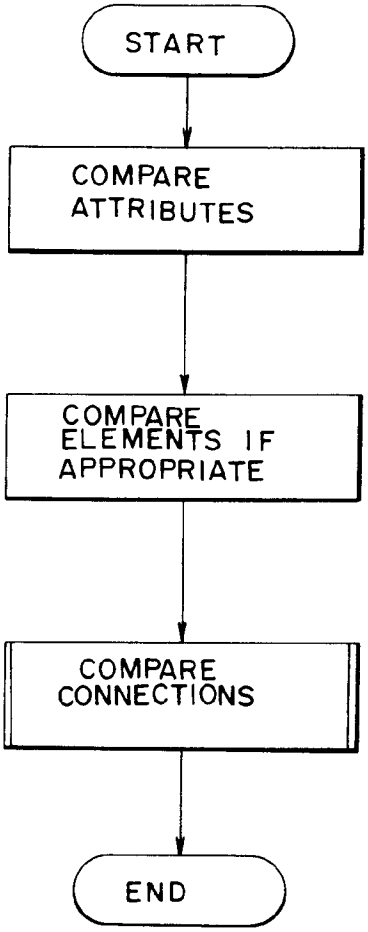

START
COMPARE ATTRIBUTES
COMPARE ELEMENTS IF APPROPRIATE
COMPARE CONNECTIONS
END

FIG. 3

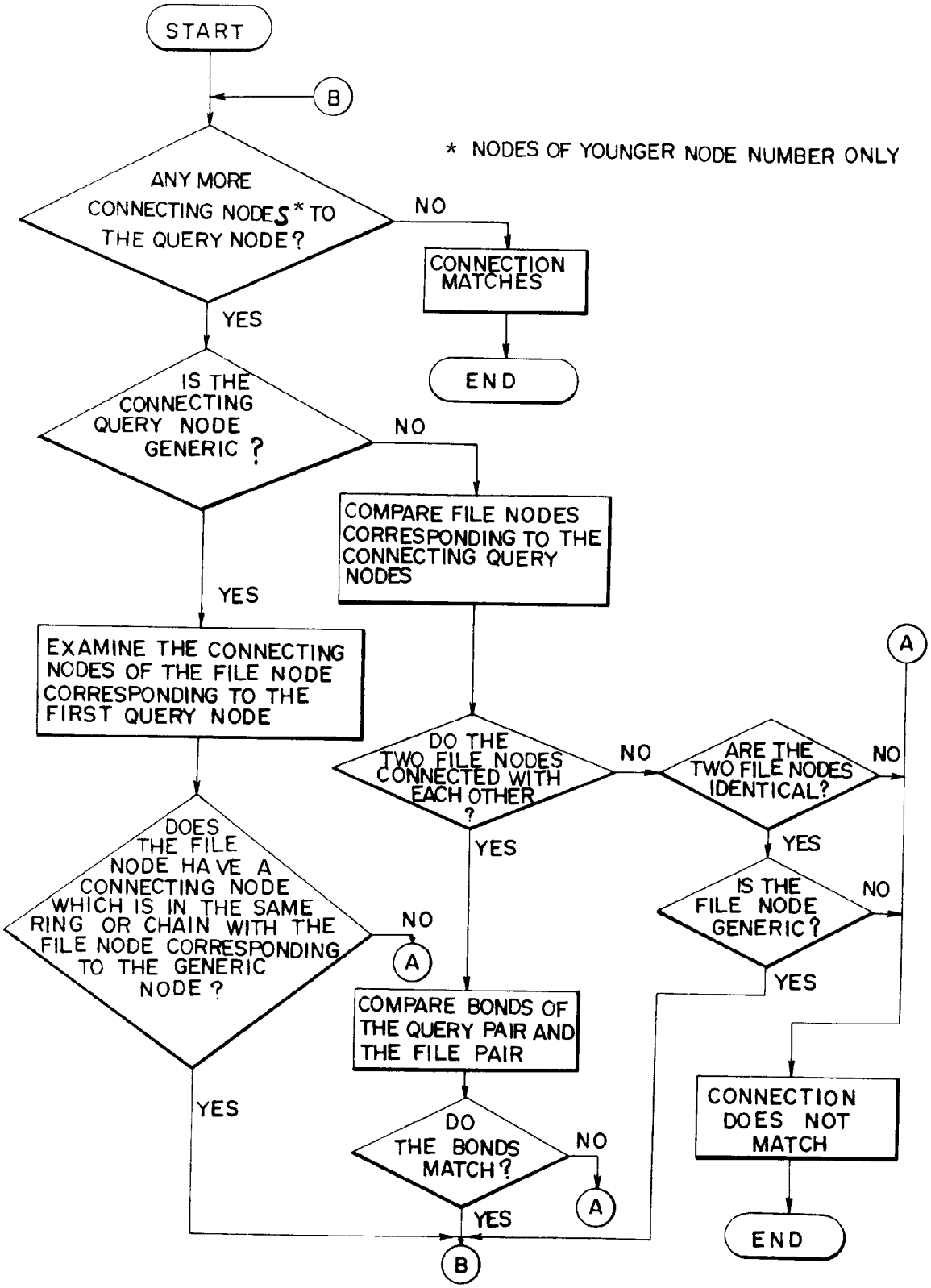

START
B
ANY MORE CONNECTING NODES* TO THE QUERY NODE?
NO
CONNECTION MATCHES
END
YES
IS THE CONNECTING QUERY NODE GENERIC ?
NO
COMPARE FILE NODES CORRESPONDING TO THE CONNECTING QUERY NODES
YES
EXAMINE THE CONNECTING NODES OF THE FILE NODE CORRESPONDING TO THE FIRST QUERY NODE
DO THE TWO FILE NODES CONNECTED WITH EACH OTHER ?
NO
ARE THE TWO FILE NODES IDENTICAL?
NO
A
YES
YES
DOES THE FILE NODE HAVE A CONNECTING NODE WHICH IS IN THE SAME RING OR CHAIN WITH THE FILE NODE CORRESPONDING TO THE GENERIC NODE ?
NO
A
IS THE FILE NODE GENERIC?
NO
YES
COMPARE BONDS OF THE QUERY PAIR AND THE FILE PAIR
YES
DO THE BONDS MATCH?
NO
A
YES
CONNECTION DOES NOT MATCH
END
B
* NODES OF YOUNGER NODE NUMBER ONLY

FIG.4

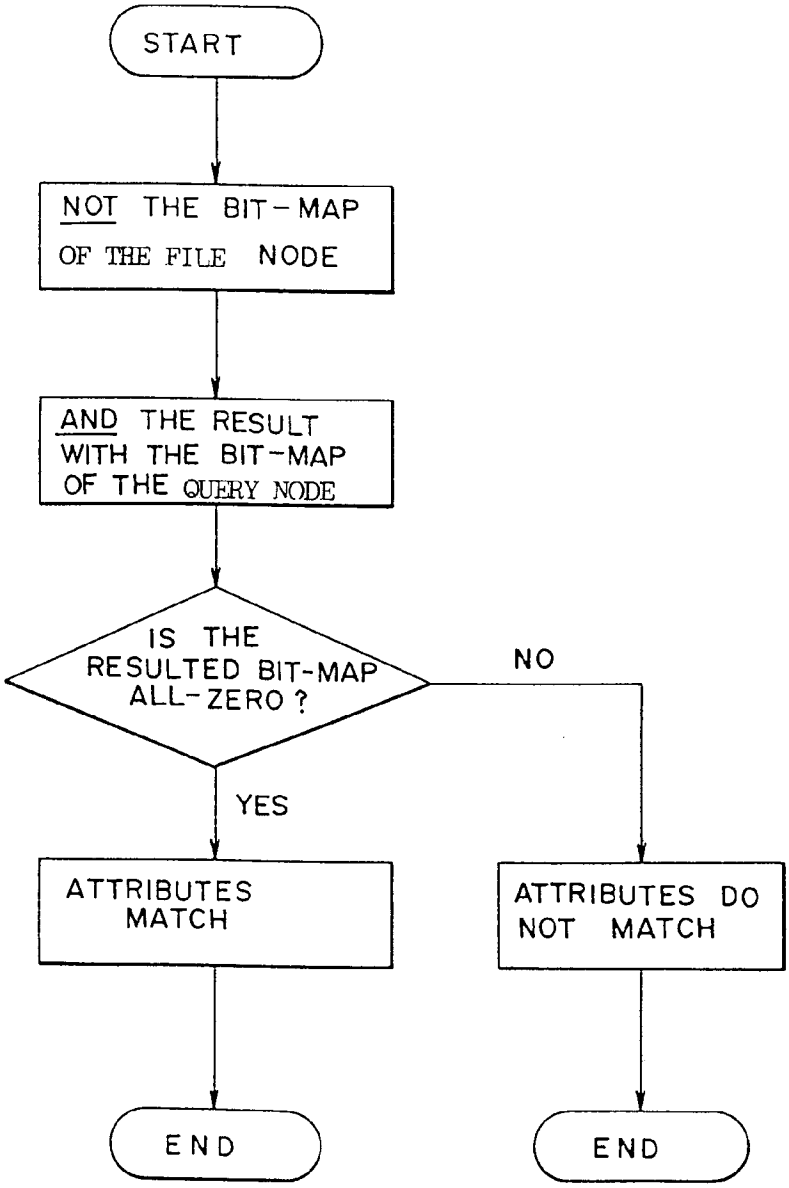

START
NOT THE BIT-MAP OF THE FILE NODE
AND THE RESULT WITH THE BIT-MAP OF THE QUERY NODE
IS THE RESULTED BIT-MAP ALL-ZERO ?
NO
YES
ATTRIBUTES MATCH
ATTRIBUTES DO NOT MATCH
END
END

FIG. 5A

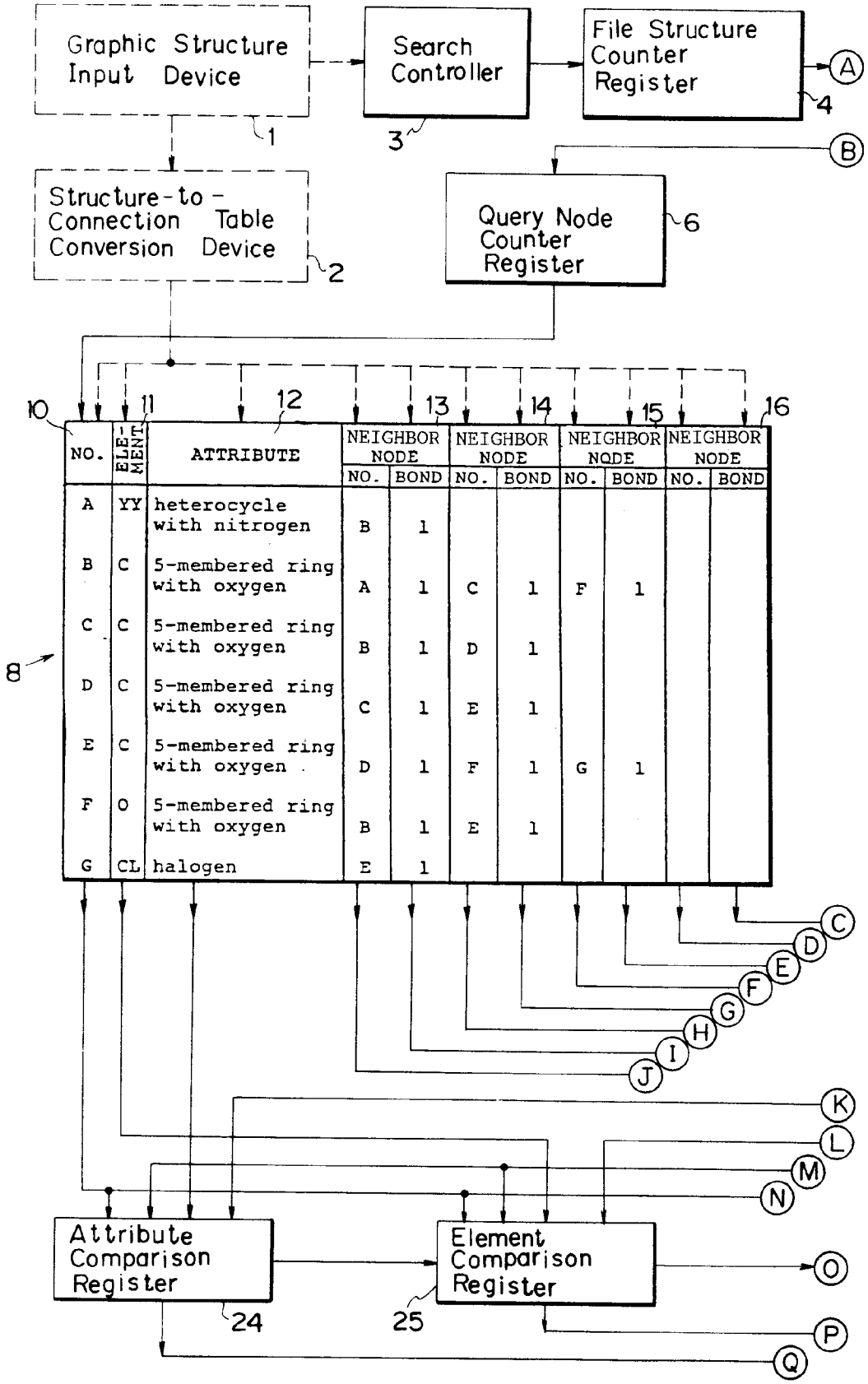
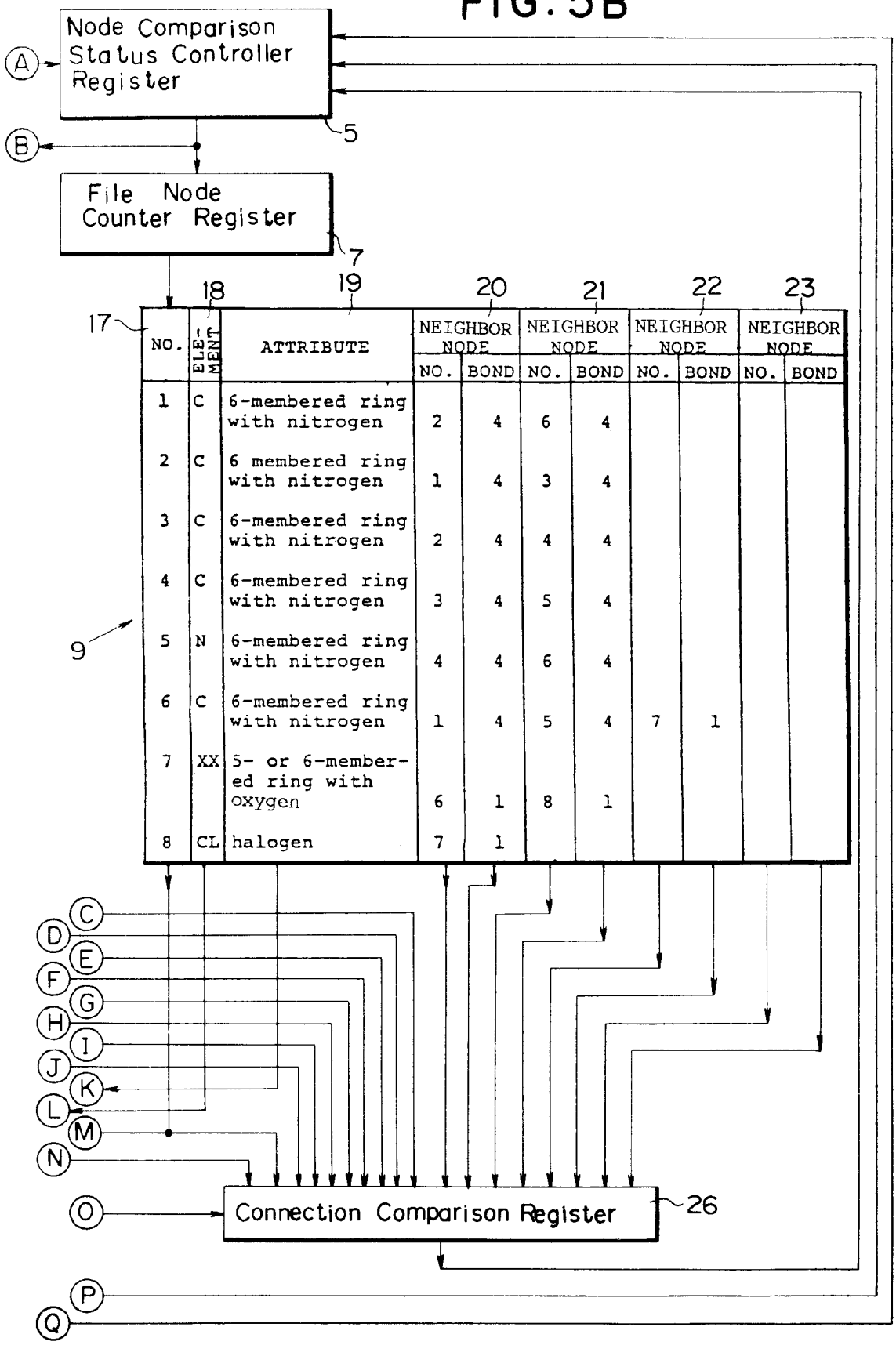

Graphic Structure Input Device
1
Search Controller
3
File Structure Counter Register
4
A
B
Structure-to-Connection Table Conversion Device
2
Query Node Counter Register
6
8
10
11
12
13
14
15
16
NO.
ELEMENT
ATTRIBUTE
NEIGHBOR NODE
NO.
BOND
A YY heterocycle with nitrogen B 1
B C 5-membered ring with oxygen A 1 C 1 F 1
C C 5-membered ring with oxygen B 1 D 1
D C 5-membered ring with oxygen C 1 E 1
E C 5-membered ring with oxygen D 1 F 1 G 1
F O 5-membered ring with oxygen B 1 E 1
G CL halogen E 1
C
D
E
F
G
H
I
J
K
L
M
N
Attribute Comparison Register
24
Element Comparison Register
25
O
P
Q

FIG. 5B


Node Comparison Status Controller Register
5
File Node Counter Register
7
17
18
19
20
21
22
23
9
NO.
ELEMENT
ATTRIBUTE
NEIGHBOR NODE
NO.
BOND
1 C 6-membered ring with nitrogen 2 4 6 4
2 C 6 membered ring with nitrogen 1 4 3 4
3 C 6-membered ring with nitrogen 2 4 4 4
4 C 6-membered ring with nitrogen 3 4 5 4
5 N 6-membered ring with nitrogen 4 4 6 4
6 C 6-membered ring with nitrogen 1 4 5 4 7 1
7 XX 5- or 6-membered ring with oxygen 6 1 8 1
8 CL halogen 7 1
A
B
C
D
E
F
G
H
I
J
K
L
M
N
O
P
Q
Connection Comparison Register
26